**I S 451 Project Report - YouTube**

**Dataset Overview**

The dataset we analyzed was collected from Kaggle, called “YouTube Channel Performance Analytics”. It contained in-depth information about important metrics in video performance like revenue generation, video performance, viewer behavior, and audience engagement. The raw dataset contained 70 columns of variables of which we selected the most relevant independent variables (video duration, day of the week, etc.) to narrow the scope of the modeling. There were 364 records in the dataset, including videos published from 2016-2024.

**Abstract**

YouTube is a video sharing platform that serves many different stakeholders: viewers, advertisers, content creators (suppliers), and Google. It has become a hub for internet culture and has continued to grow over the years, bringing in over $31 million in advertising revenues in 2023 alone (Appendix 2). Viewers are the consumers of the videos, being the people that content creators and advertisements try to reach. They simply watch, react, and share videos, or even subscribe to their favorite content creators. Creators make videos of all types and can earn money on the platform through monetization of their videos after fulfilling certain viewer and subscriber count requirements. If they satisfy these requirements, creators can join the Youtube Partner Program (YPP) where they can earn money from their videos (Marshall). Members must have at least 1000 subscribers, 4000 watch hours, and no community strikes against your YouTube account in order to monetize their videos for views and ad impressions (Blogging Wizard). There is also a lower tier of monetization known as fan funding, which allows creators to capitalize on viewers donating money (Appendix 1). This functions similarly to tipping in the United States. YouTube separates their monetization into 3 modules: Commerce Product (channel memberships, super chat, etc.), Watch Page Monetization (ads displayed/streamed on videos), and Shorts (share of views in Creator Pool allocation). Creators can earn 70%, 55%, and 45%, on net revenues allocated, under these modules, respectively. Our group was curious about what factors (time, video duration, likes, etc.) contribute the most to YouTube and video revenues. This would be helpful for Google, content creators, as well as numerous curious parties wondering how much specific videos and content creators make.

**Our Goal & Stakeholders**

There are two main stakeholders we considered in our project: content creators and YouTube/Google. Our main goal was to create a model that could reasonably predict the amount of revenues based on such variables as Day of Week, Video Publish Time, Views Shares, Video Duration, Dislikes, Likes, and Subscribers. Of the seventy potential variables offered in the dataset, these stood out to be the ones that would be reasonable predictors to consumer engagement, which ultimately dictates content revenue. This dataset does not concern video ads where content creators cultivate their own revenue through personal brand deals.

There were several problems we encountered when coding. The first one was the limited size of the dataset with 364 rows. It is too small to sufficiently train a machine learning model. However, to address this, we implemented k-fold cross-validation, a technique that maximizes the utility of limited data by dividing it into multiple subsets for training and validation. This approach not only ensured that each data point was utilized in both training and validation but also enhanced the model’s generalization by reducing the risk of overfitting. By evaluating the model’s performance across different folds, we were able to achieve a more reliable estimate of its effectiveness, despite the constraints of a small dataset. Another challenge we encountered during the project was the varying scales of the features in our dataset. For example, when visualizing the data using techniques like boxplots and histograms, we observed that some features, such as "views," ranged from 0 to 700,000, while others, like "video duration" and "subscribers," had much smaller ranges, typically between 0 and 4,000. This disparity in scale posed a risk of certain features dominating the model and skewing the results. To address this, we applied normalization techniques to rescale all features to a common scale, ensuring that each feature contributed equally to the analysis and model training. In addition, based on the visualization of all features, we discovered that there are some outliers in both independent and dependent variables. Typically, we would drop the outliers, but due to the limited data, we created different models with and without outliers to compare which model had better performance based on metrics.

**Data Cleaning & Adjustments** (See Exhibit 1)

We began analysis with an initial data inspection using data.info() and data.duplicated() to check for null values, data types, and duplicates. The Publish Time column was then split into separate columns for Hour, Day, Month, and Year, and we dropped the original column. The Days column was converted to integers (1-7) to represent the days of the week. We identified negative values where it did not make sense, such as a value of -3 in the Subscribers column. These anomalies were addressed by dropping the respective row that contained the negative values. We did not want the illogical values to potentially skew the dataset, especially with a limited amount of data entries. Also, we found that Duration, Views, Revenue, Subscribers, Shares, and Likes were concentrated on the lower end, likely because only a small percentage of videos achieve significant success. Additionally, we noted that 2017 had a much higher number of published videos compared to other years so we allocated different sample weights based on the ratio of existing data when we split the dataset.

**Methodology**

Of the seventy columns, we wanted to focus on the variables that were relevant to estimating video revenues. We narrowed down the independent variables to eight—Day of Week, Video Publish Time, Views, Shares, Video Duration, Dislikes, Likes, and Subscribers. In addition, due to the limited rows, we used the K-fold cross-validation technique to maximize the allocation of our data, which improved the generalizability of our results/models.

We generated two models with our dataset: 1) a model to predict revenues and 2) whether or not the revenue of a given video is greater than the median of all revenues. At first, we tested four linear models with different parameters—Random Forest, Gradient Boosting Regressor, and Support Vector Regressor model. Out of all 4 models, we selected the Random Forest model because the R-squared value was 0.53 and the MSE of the model was about 17, which was acceptable due to the range of estimated revenue (from 0 to 103) and had a higher metrics score than other linear models. Overall, the linear models had pretty high MSE values and low R-squared values, meaning the model was not a great fit for the data. However, we still wanted to keep a way to predict revenues, so we kept the linear regression model for our data.

To supplement the linear model, we created another model as a binary model to predict whether the estimated revenue would be greater than the median revenue or not (1/0). We decided not to use the average revenues as they are very skewed towards those who make a lot on YouTube. The Gradient Boosting Classifier model had higher performance scores than the other binary models Random Forest Classifier and the Support Vector Classifier, with a 0.83 accuracy score, 0.76 precision score, 0.88 recall score, and 0.81 F-1 score.

**Further Modeling**

We also used other methods to gain more insights on the data we were working with. Box plots were used to visualize the data and any outliers. From the box plots of the different independent variables, we saw that there were numerous outliers. These points skew the data, so when moving forward with our methodology, we created models with and without outliers (See Exhibit 2). We then created scatterplots to show the relationship of key variables (Likes, Subscribers, Shares) with revenue (See Exhibit 4). Each scatterplot showed positive correlation between the independent and dependent variable.

We used the correlation matrix (See Exhibit 3) to visualize the correlation between specific variables. Estimated Revenue was most heavily correlated with Views, Shares, Likes, and Subscribers with correlations of 0.36, 0.36, 0.43, and 0.42 respectively. This makes sense as these metrics are the typical measures of what we consider to make a YouTube video successful. We found in this matrix that Day of Week had the weakest correlation and even had a negative value of -0.06. Video Duration and dislikes surprisingly had minimal correlation with revenues at 0.14. We thought it may have a higher correlation because a longer video would allow creators to satisfy video duration requirements for specific adverts—mid-roll ads become possible after reaching the ten-minute threshold (Davis).

Dislikes are an interesting metric to follow because you would think it would be negatively correlated with revenues, but in the case of media or entertainment there is the saying that “any publicity is good publicity”. There are certainly times where dislikes could be an indicator of the virality of a video. It is common in the mass media era for videos or creators to gain massive amounts of attention that are mostly fueled from negative online energy. KSI is a perfect example of this where his recent release “Thick of It” received heavy backlash and attention online for being a “bad song”. Due to its viral nature, it climbed to number one on the US Spotify viral chart despite it being unpopular (Alston). It became an internet meme to listen to the song and give a negative opinion on it in a humorous fashion.

**Results**

Using Models without outliers, we found there to be notable performance differences in the algorithms tested. The Random Forest achieved the best overall performance with an MSE of 16.97 and a R-Square score of 0.53. These values indicate about a 53% variance in the target variable. Gradient Boosting Regressor also performed well, with an MSE of 18.99 and an R-Square score of 0.475, demonstrating moderate predictive power when put to the test. Linear Regression, while simpler, showed limited efficacy, with an MSE of 29.81 and an R-Square score of only 0.176, indicating it captured a mere 23% of the variance. Lastly, the Support Vector Regressor (SVR) displayed the weakest performance, with an MSE of 39.40 and an R-Square score of -0.09, suggesting it was not well-suited to this dataset. From these results, we found that the Random Forest had the strongest predictive power with the dataset at 53% efficacy. The Random Forest algorithm is more apt in handling nonlinear relationships, which best matches our data (See Exhibit 5).

When running the Logistic Regression without outliers we also saw notable variation here in the algorithms tested (See Exhibit 6). The Gradient Boosting Classifier achieved one of the highest overall performances, achieving an accuracy of 83.1%, precision of 75.9%, recall of 88%, and an F1 score of 81.5%, suggesting a similarly effective performance to Random Forest, albeit higher performing in all categories, but Recall. It was better predicting the true positives than the ratio of all the positives predicted being correct. The Random Forest also performed quite well, with an accuracy of 81.4%, precision of 71.9%, recall of 92%, and an F1 score of 80.7%. This indicates that it performed reasonably well in identifying positive cases and minimizing false positives. The Support Vector Classifier (SVC) demonstrated a moderately successful performance, with an accuracy of 64.4%, precision of 57.7%, recall of 60%, and an F1 score of 58.8%. The SVC found it more difficult to maintain high success rates in its prediction. Logistic Regression exhibited the weakest performance, with an accuracy of 57.6%, precision of 50%, recall of 48%, and F1 scores of 49%, reflecting the most limited ability to predict in the dataset. After reviewing the cumulative results we found the Gradient Boosting Classifier to be the most effective model for this dataset as a whole.

With the Gradient Boosting Classifier model being the best fit for the data, we investigated which independent variables were the most important when making a predictive decision. The most important features of the model were the Publish Year and Likes, and with scores of 0.192297 and 0.131159 respectively. Subscribers, Dislikes, Shares, Video Duration, and Views were all moderately important with scores ranging from .09-.12. All other independent variables (Days, Publish Day, Publish Month, and Publish Hours) showed relatively low importance to the model. We would also like to note that the Publish Hour was a non-factor in this model as it showed no importance with a 0.0 importance score. Although having low importance to the Gradient Boosting Classifier it is an interesting note that the Day\_2 column (Tuesday) has a 0.012 importance with the next closest day being Day\_3 (Wednesday) at 0.008.

**Discussion & Conclusion**

Prior to our data analysis, we wondered whether or not there was any correlation between the timing of when content creators would post and their revenues. Using the Gradient Boosting Classifier, we found that the specific time in terms of month, day, and hour were not very important. There may have been some times that could have a slight impact on engagement and success, but what mattered the most was the Publish Year and Likes. Content Creators do not need to be too concerned with when they post their videos, as posting a video based on the year is not very reasonable. Much of that difference coming from different Publish Year is more reflective of the growth and success of YouTube as a whole than any conscious decision made from a content creator.

Furthermore, we found that YouTube was most successful in 2022 in terms of revenues (Exhibit 9), which was surprising because we expected it to be the highest in 2020 when people were at home due to the pandemic. There are several possible reasons why YouTube was most successful in 2022. One of the reasons was because YouTube heavily promoted Shorts that year to compete with TikTok and Reels. It was also ranked #2 most popular social networks in 2022, with 2.5B active users. According to Statista, viewers watched YouTube for 19 hours weekly, compared to 18 from 2021 and 17 hours in 2023. Because creators don’t have control over what year YouTube is the most successful, we recommend that they focus mostly on what kind of audience they want to bring in and target their content towards them to garner more likes, and in turn, revenues. There were no concrete success stories for YouTube’s success in 2022 aside from Shorts, but not enough data to recommend creators focus on making shorts. As for recommendations for YouTube, we recommend that they promote their Shorts like they did in 2022, as this was the only factor we found online that could have contributed to their high median revenues in that year. As YouTube’s goals align with content creators, they should continually improve algorithms to push relevant and interesting content to their viewers.

We wish our data had more demographics within it such as age group, country, and gender of viewers, as well as video descriptionsto see what types of videos appeal to different groups of people. This could open the data up for analysis amongst stakeholders like advertisers (who to target and on what videos) and viewers, and also give more useful insights to content creators so they know what type of videos are popular amongst certain demographics. We also wish our data had more points in general, as 364 records is not nearly sufficient enough to train a model. More data points would strengthen the importance scores and allow us to better generalize our findings to give more concrete recommendations and analyses.

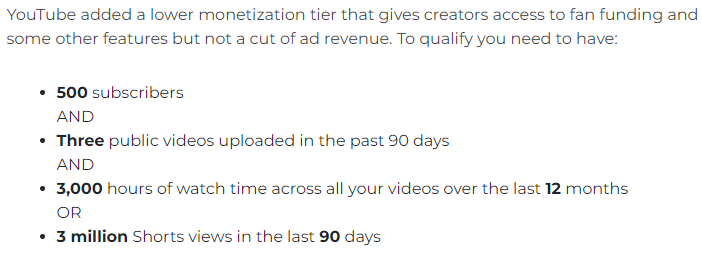
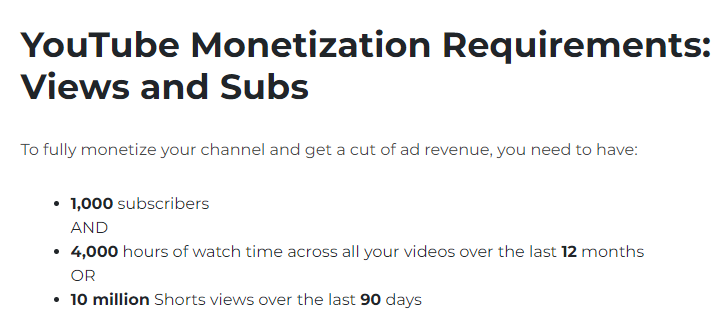
As for how we would improve our own data / do things differently, there were negative numbers of subscribers for certain records, which shouldn’t be possible. We would want to correct those numbers instead of removing it from the dataset. We could also adjust the revenues per year for inflation. Furthermore, as we did our model without outliers, we would want to validate these data points and check if they are real. Once we validate them, we could rerun the models with outliers, because some creators just make that much more money from their videos. Finally, we might retrain the model without the Publish Day of the month (for example, 17th of May, or days 1-31), to reduce the possibility of collinearity with the day of the week, which is the more important variable (Monday-Sunday).

Overall, with the data that was available to us, we were able to find some interesting points, namely Publish year being the most important factor in predicting revenues. A more extensive dataset would definitely strengthen our arguments and our recommendations.

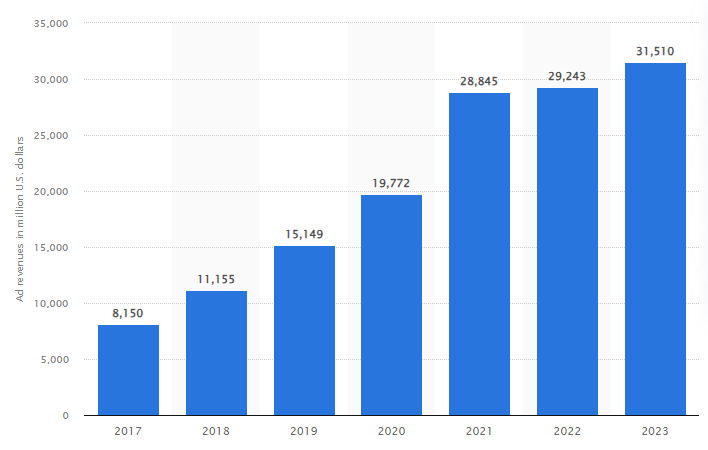
.

**Appendix**

1. **(Marshall)**



1. **(Ceci)**

****

**Exhibits**

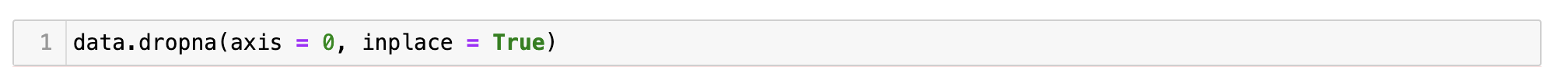
Exhibit 1: Data Cleaning

Data Cleaning:

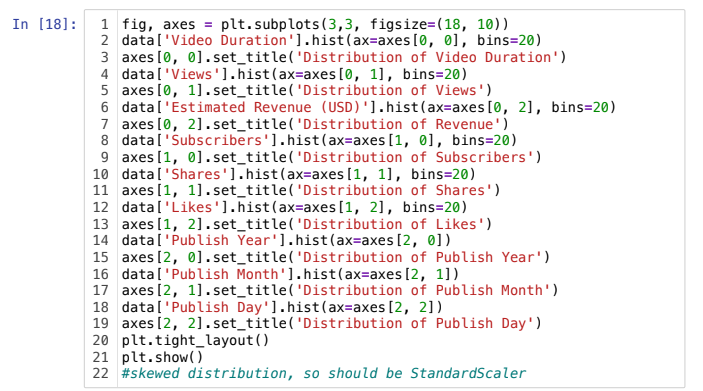
* Used the data.info() and data.duplicated() function to check if there were null values, data types, and duplicate values / nonvalues.
* Publish Time was changed to Hour, Day, Month, and Year in their own columns. After this was accomplished, the Video Publish Time column was dropped from the dataset.
* The “Days” column was converted to integers 1-7 to represent Monday through Sunday.
* We found there to be some negative values in columns where it would not make sense. For example, in the Subscribers column there were values of -3. This was discovered after analyzing the data with data.describe() and checking if the Min made sense. We noticed the negative anomaly and wanted to adjust such inputs to not impact the dataset as drastically, see below. Having a negative number of subscribers to a channel is not possible, so we dropped the rows that contained such values. This was done because the dataset size is not that big and the column is skewed.







* Breaking down the variables, we took a look at each distribution model to see if there was an uneven concentration of data inputs which could skew the data one way or another. For Duration, Views, Revenue, Subscribers, Shares, and Likes most of the data was on the lower end. A possible reason for this occurrence is due to the percentage of videos that end up being “successful” is far lower than the total volume of videos. Many videos may be posted, but only a few of them return a significant amount of views and viewer interaction. For the Publish Year, there were far more videos around 2017 than any other year, which is an important note when analyzing end results.



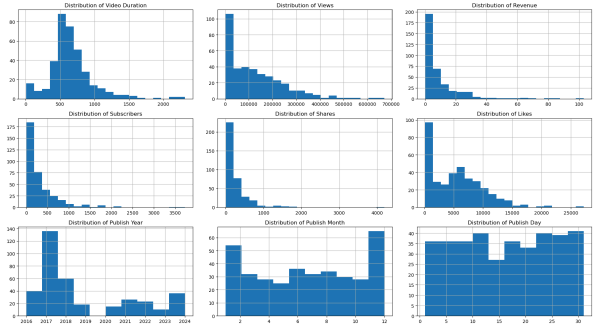
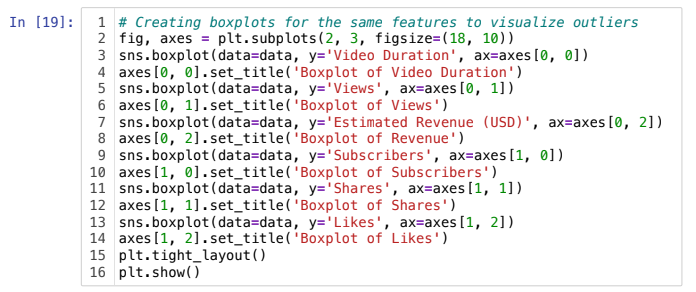


Exhibit 2: Box Plots



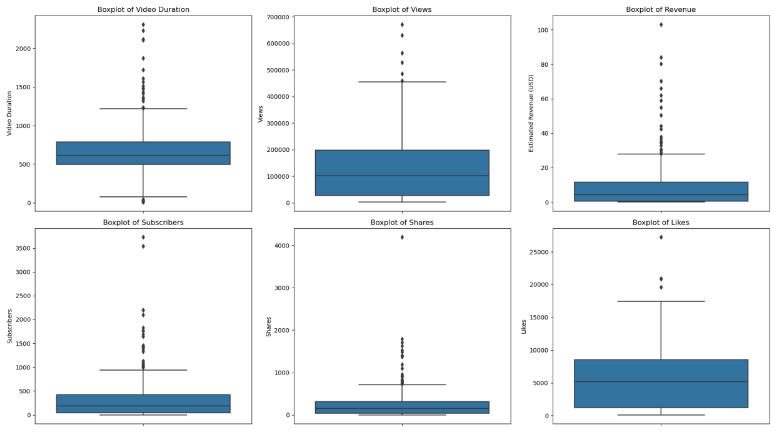
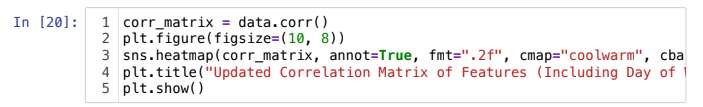


Exhibit 3: Correlation Matrix



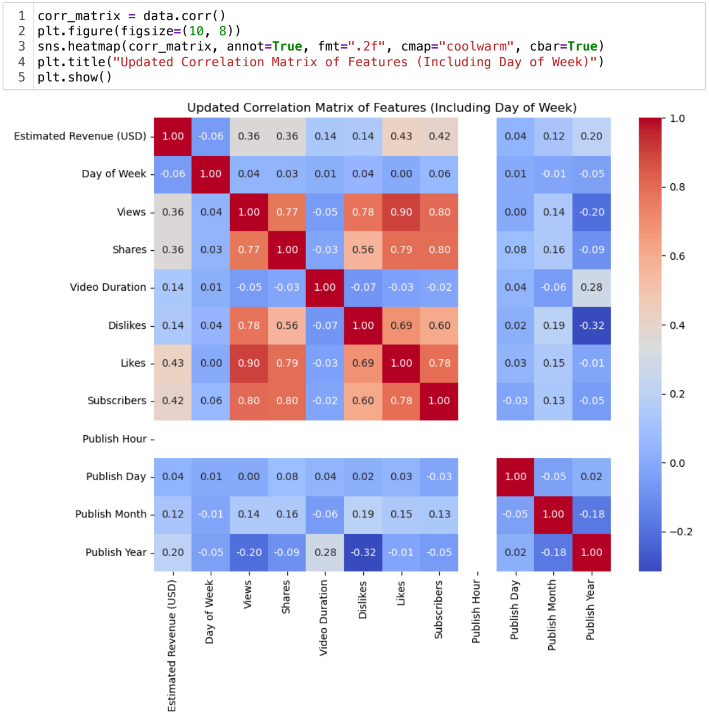
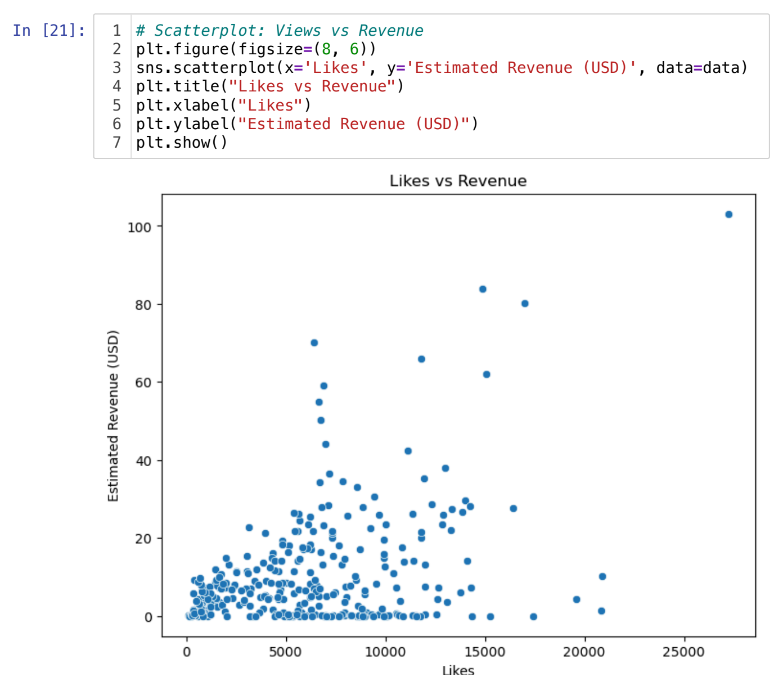
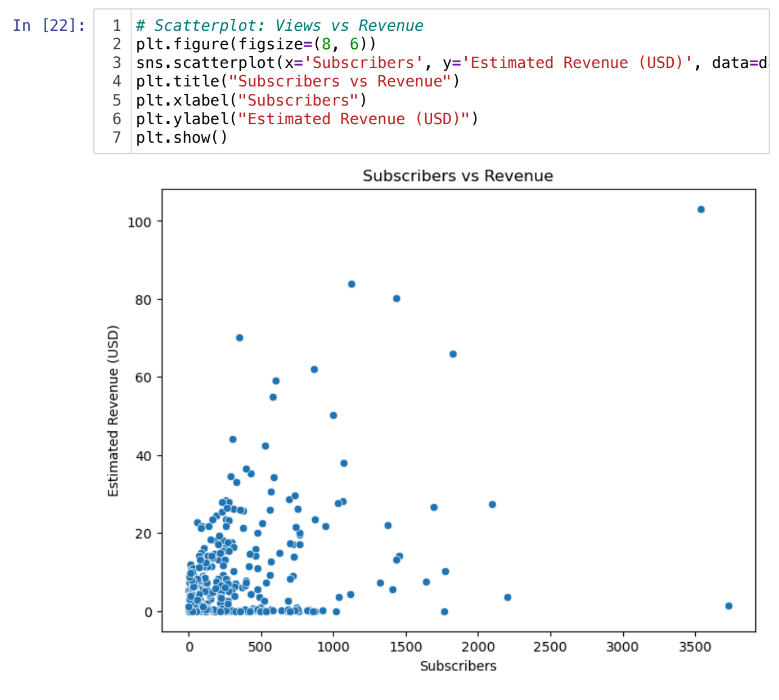


Exhibit 4: Scatterplots





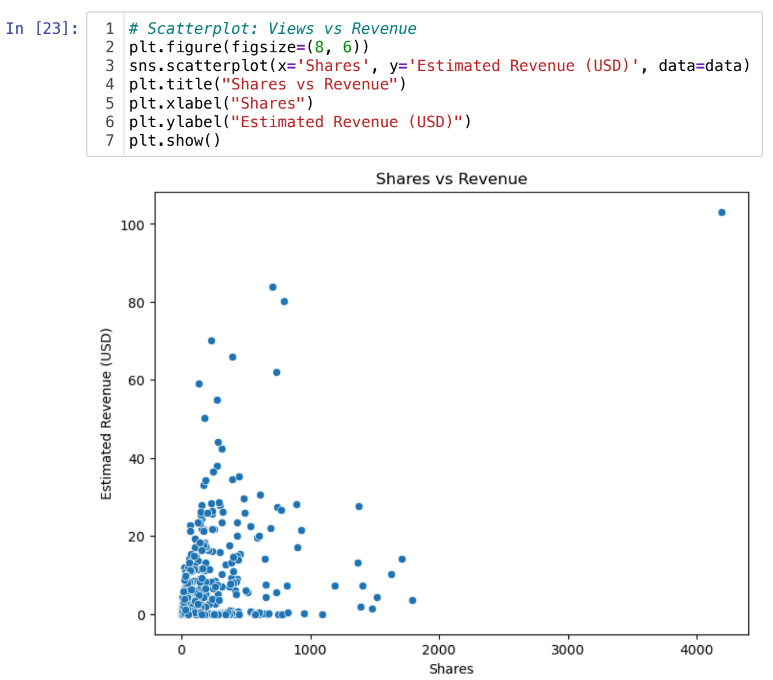
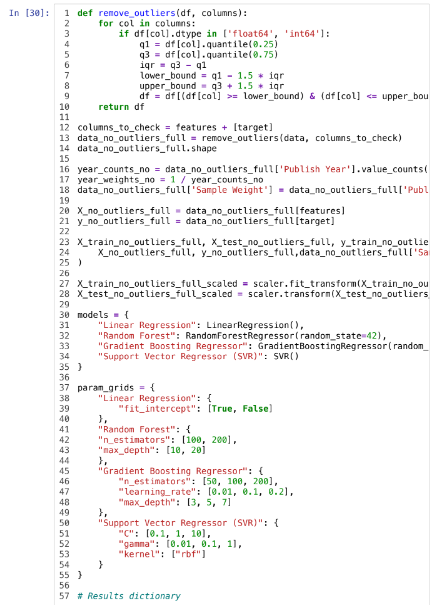
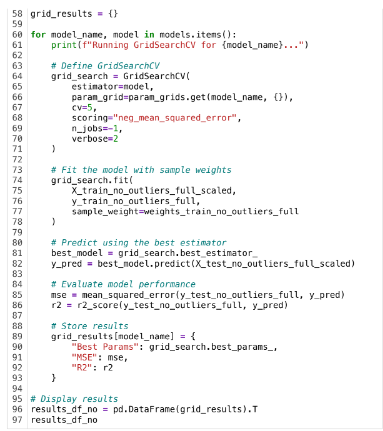


Exhibit 5: Linear Model without Outliers





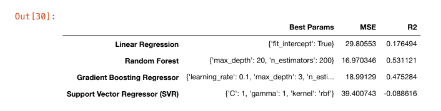
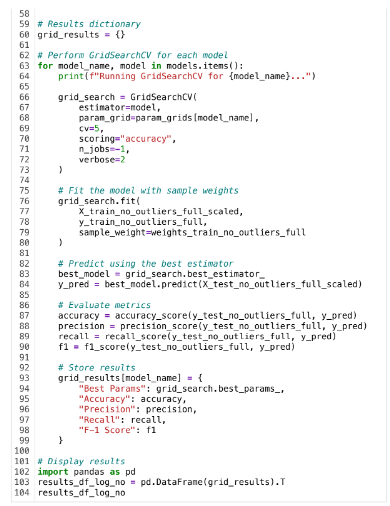


Exhibit 6: Binary Model without Outliers





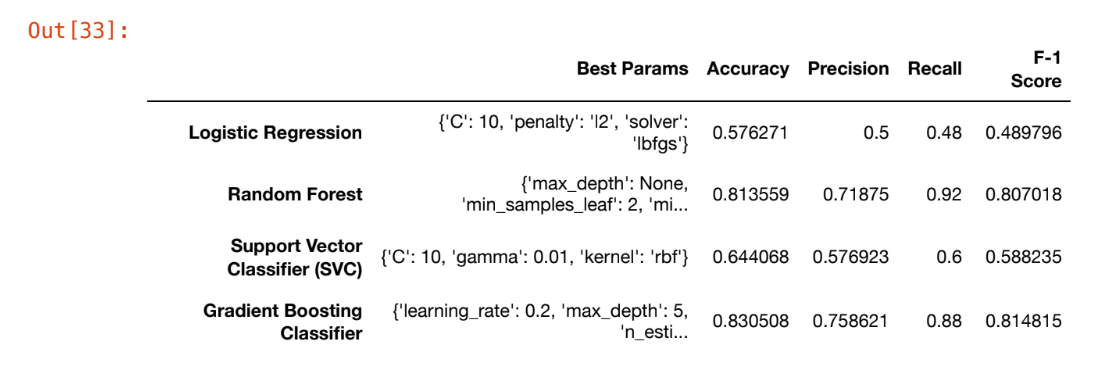
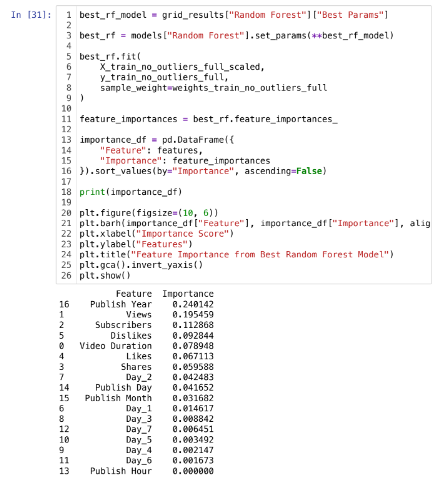


Exhibit 7: Variable Importance for Random Forest Model



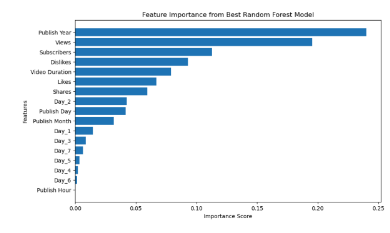


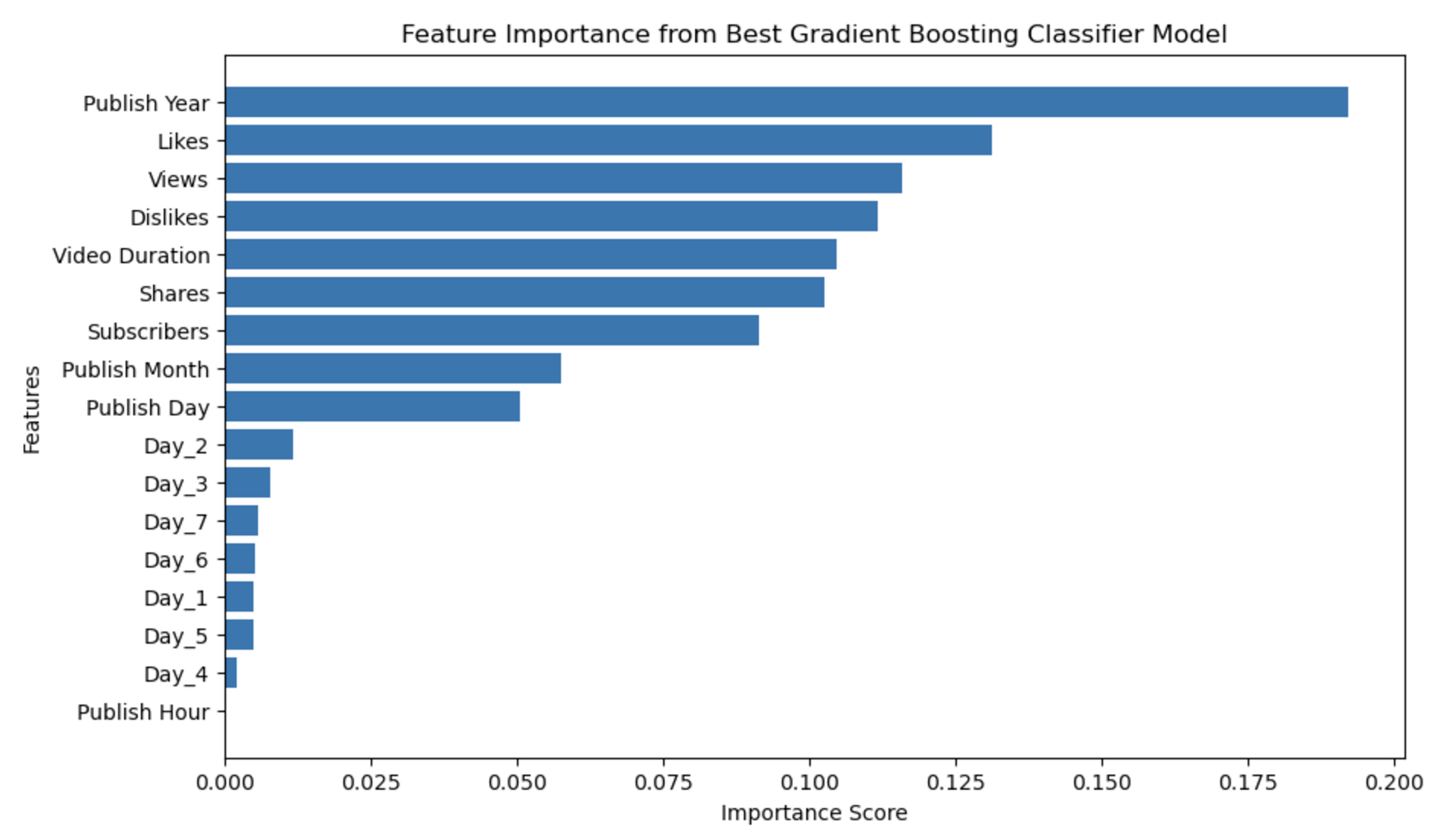
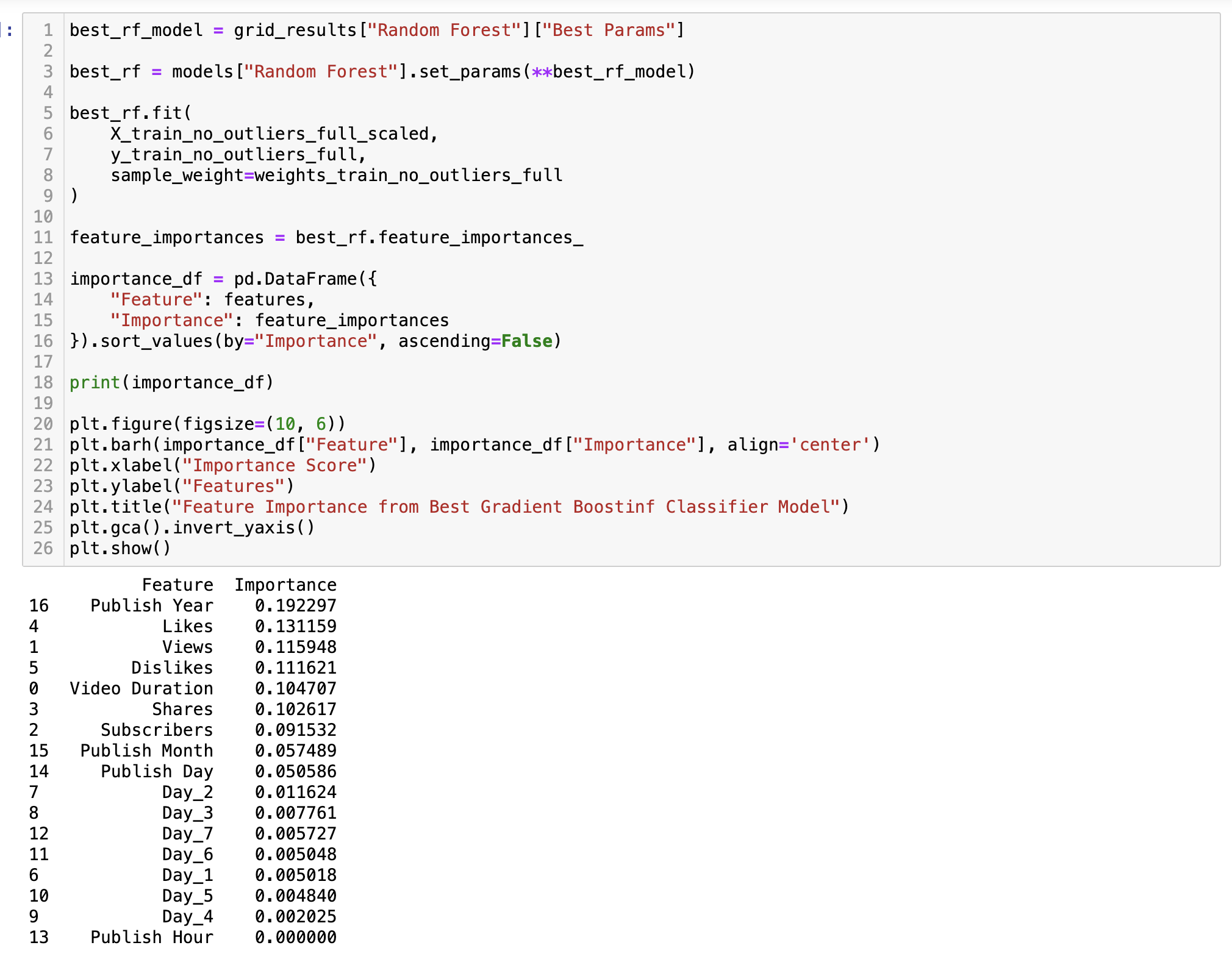
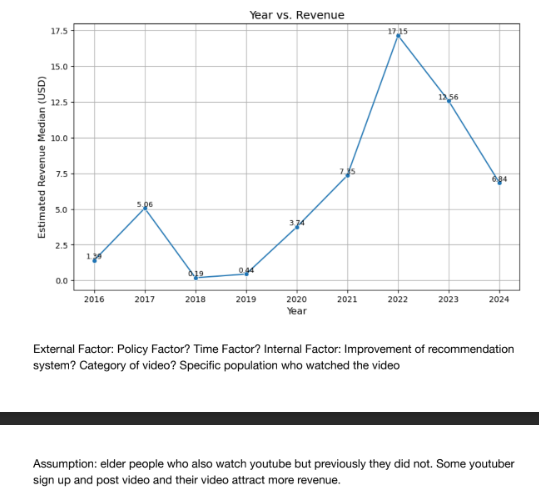
Exhibit 8: Variable Importance for Gradient Boosting Classifier Model

Exhibit 9: Estimated Revenues and Year for YouTube



**Works Cited**

Alston, T. (2024, October 22). Despite fans hating KSI’s song “Thick of It,” it’s climbing the charts. *Complex*.<https://www.yahoo.com/entertainment/despite-fans-hating-ksis-song-052718429.html>

Blogging Wizard. (n.d.). How many YouTube subscribers do you need to make money?<https://bloggingwizard.com/how-many-youtube-subscribers-to-make-money/#:~:text=YouTube%20requires%20members%20to%20have,play%20during%20your%20YouTube%20videos>.

Ceci, L., & 21, M. (2024, May 21). YouTube global advertising revenues 2023. *Statista*.<https://www.statista.com/statistics/289658/youtube-global-net-advertising-revenues/>

Comedy Central. (2012, October 17). *Key & Peele - Substitute Teacher* [Video]. YouTube.<https://www.youtube.com/watch?v=Dd7FixvoKBw>

Davis, J. (2022, January 25). Video length for YouTube monetization (What you need to know). *Society of Side Hustle-Go to Side Hustle Site*.<https://societyofsidehustle.com/video-length-youtube-monetization/>

Google. (n.d.). What's my revenue share? *YouTube Help*.<https://support.google.com/youtube/answer/72902?hl=en#zippy=%2Cwhats-my-revenue-share>

Kaggle. (n.d.). YouTube channel performance analytics.<https://www.kaggle.com/datasets/positivealexey/youtube-channel-performance-analytics>

Marshall, C. (2024, September 26). YouTube monetization requirements: What you need to know in 2024. *TubeBuddy*.<https://www.tubebuddy.com/blog/youtube-monetization-requirements/>

Office Timeline. (n.d.). YouTube history timeline.<https://www.officetimeline.com/blog/youtube-history-timeline>

Soax. (n.d.). Top social media platforms: Statistics 2024.<https://soax.com/research/top-social-media-platforms#:~:text=Research%20highlights%3A%20In%202024%2C%20Facebook,each%20with%202%20billion%20users>.

Statista. (n.d.). Global social networks ranked by number of users.<https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>

Statista. (n.d.). Online video time spent.<https://www.statista.com/statistics/611707/online-video-time-spent/>

Thought Leaders. (2022). 2022 in review: YouTube edition.<https://www.thoughtleaders.io/blog/2022-in-review-youtube-edition>

YT Large. (n.d.). YouTube video data viewer.<https://ytlarge.com/youtube/video-data-viewer/>

YouTubers.me. (n.d.). Nigahiga YouTube videos stats.<https://us.youtubers.me/nigahiga/youtube-videos-stats>